

WP. 18
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (iii): Research data centres and virtual labs

**UK SECURE DATA SERVICE: SPECIFICATIONS AND CHALLENGES OF USING
POTENTIALLY DISCLOSIVE DATA**

Invited Paper

Prepared by Reza Afkhami, Melanie Wright and Mus Ahmet, UKDA, University of Essex, United Kingdom

UK Secure Data Service: specifications and challenges of using potentially disclosive data

Reza Afkhami, Melanie Wright, Mus Ahmet
UKDA University of Essex
rafkhami@essex.ac.uk

Abstract

The Secure Data Service is a secure environment funded by ESRC to provide researcher access to disclosive micro data either from their offices, safe rooms in their institutions or on site at the UKDA. Its operation is legally framed by the 2007 statistics Act which makes access to the confidential data for statistical purposes possible. This short paper introduces this new UKDA service with its proposed specifications and also some challenges facing data service providers. It is envisaged that the proposed SDS infrastructure will be able to meet the requirements of the data security model.

1. Introduction

Disclosure of personal information can be harmful. This may result in being denied for civil services, embarrassment and loss of reputation and trust. Less directly, research results based on disclosive data can cause harm by affecting perceptions about a group to which a person belongs.

A key problem in Secure Data Service (SDS) data confidentiality is to balance the legitimate requirements of data users and confidentiality protection. The confidentiality of individual information can be protected by restricting the amount of information provided i.e. by adjusting the data in released tables and statistical outputs (restricted data) or by imposing conditions on access to the data products (restricted access), or by some combination of these.

The UKDA Secure Data Service is a new service to allow controlled restricted access procedures for making more detailed microdata files available to some users (Approved Researchers), subject to conditions of eligibility, purpose of use, security procedures, and other features associated with access to the SDS data.

This short paper will introduce this new UKDA service with its proposed specifications and also some challenges facing data service providers. It is envisaged that the proposed SDS infrastructure will be able to meet the requirements of the data security model.

Building on the success of other secure data enclaves worldwide, and employing security technologies used by the military and banking sectors, the SDS will allow trained researchers to remotely access data which is held securely on central SDS servers at the UK Data Archive. The aim of the service is to provide approved academics unprecedented access to valuable data for research from their home institutions, with all of the necessary safeguards to ensure that data is held, accessed and handled securely.

The SDS follows a model which suggests that the safe use of data should cover the elements of safe project, safe people, safe setting and safe output (Ritchie, 2006. see Figure 1). In order to achieve the above goal, data security depends on a matrix of factors, including technical, legal, contractual, and educational. The structure of this paper revolves around these factors demonstrating how SDS has set up the necessary infrastructure to meet these requirements.

In doing so, we discuss the technical features and the system specifications followed by the legal and contractual responsibilities of the users and their institutions. The issues like users' education and training before granting access to data are examined. We also talk discuss the kind of disclosive data we are aiming to support in the SDS. Finally, the challenges facing the SDS operation will be examined.

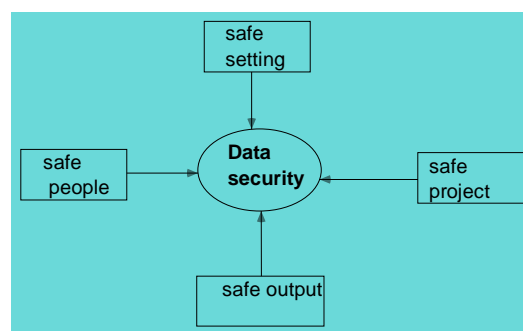


Figure 1: Elements of data security

2. System Specifications

The technology used by the SDS must be secure and the system adheres to the highest and most transparent quality standards. The technical model that has emerged is one which shares many similarities with both the ONS VML¹ and the NORC Secure Data Enclave². It is based around a Citrix infrastructure which turns the end user's computer into a 'remote terminal' giving access to data, statistical software, and collaboratory spaces on a central secure server held within the UK Data Archive. The system is flexible, in that depending upon the wishes of the data custodians, access can be restricted to particular users (safe people) and/or particular locations (safe rooms/machines). It is secure because all data manipulation occurs on the server, which is maintained to very strict security protocols.

Beyond the general security policy, the secure server itself will be subject to additional security measures and controls. Approved researchers will access the proposed SDS by using VPN (Virtual Private Network/thin-client) technology, which encrypts the data transmitted between the researcher's computer and the host network. Other components of the VPN technology allow control to be established over which network resources the external researcher can access on the host network. The service will employ a Citrix XenApp server farm, which participates on two networks.

¹ <http://www.ons.gov.uk/about/who-we-are/our-services/unpublished-data/business-data/vml/index.html>

² <http://www.norc.org/DataEnclave>

How the system operates?

With this technology, although all applications (SPSS, STATA, etc) and data run on a central server at the UKDA/SDS, the Approved Researcher still interacts with a full Windows graphical user interface. This means that the researcher never has to install any complex applications on his/her remote computer - the only application required by the Approved Researcher is a web browser. This also means that the UKDA can prevent the researcher from transferring any data from the data archive to a local computer. For example, Citrix can be configured so that data files cannot be downloaded from the remote server to the user's local PC. Similarly, the Approved Researcher cannot use the "cut and paste" feature in Windows to move data from the Citrix session into an Excel spreadsheet sitting on the local computer. Finally, the user is prevented from printing the data on a local computer. The Approved Researcher logs onto the SDS system remotely via a web secure (HTTPS) browser. All data processing is carried out on a central secure server, which processes all requests centrally and returns information about the results. No data travels over the network, except the statistical results sent from the central server to the remote location by an encrypted email after the final outputs are checked against statistical disclosure controls.

Key Features

- Clients cannot remove data
- Absolutely no webpage access
- Clients cannot import data
- Data transfers are logged
- All traffic is encrypted
- Auditing
- Critical Security updates are applied daily

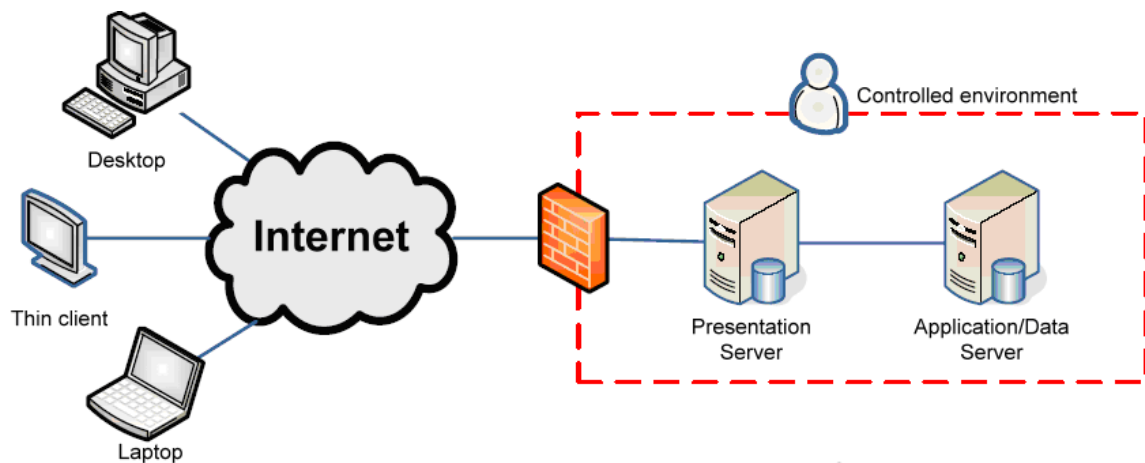


Figure 2: SDS system architecture

3. Legal and contractual framework

Users of the SDS will be required to be either “ONS Approved Researchers”³ or “ESRC Accredited Researchers”. The first of these is defined by the Statistics and Registration Services Act 2007 as “an individual to whom the Board has granted access, for the purposes of statistical research, to personal information held by it.”⁴ There is currently no definition of an “ESRC Accredited Researcher”, but we assume that it will have a similar status to an ONS Approved Researcher, i.e. a person who has been granted access for the purposes of statistical research to personal information which has been licensed to the ESDS/UKDA/⁵University of Essex for dissemination on behalf of a government department or some other data provider. Neither of these two types of user will be able to use the SDS without appropriate training. Mandatory training will allow the UKDA to ensure that end-users are fully aware of any penalties which they might incur if they cause a breach. We believe that if there is user approval to any penalties for breaches, and that they believe that these penalties are reasonable and necessary we will avoid the inadvertent disclosure that social science researchers are most likely to be prone to.

The 2007 Act also allows for increased sharing of data between ONS and other Departments, subject to agreement by Parliament on a case-by-case basis. At the same time the Act also outlines measures designed to protect the confidentiality of personal information. The Act states that a person who discloses personal information “is guilty of an offence and liable — (a) on conviction on indictment, to imprisonment

³ <http://www.data-archive.ac.uk/orderingData/agreements/ARFormsandNotes.doc>

⁴ Statistics and Registration Services Act 2007 § 39 (5).

⁵ <http://www.esds.ac.uk/aandp/access/licence.asp>

for a term not exceeding two years, or to a fine, or both; (b) on summary conviction, to imprisonment for a term not exceeding twelve months, or to a fine not exceeding the statutory maximum, or both.”⁶

The SDS will immediately suspend access to the service if it believes that any user is perpetrating or attempting to perpetrate any of the breaches listed in SDS security breaches or SDS confidentiality agreement. A full investigation will then follow.

Users will be required to fill out an on-line form which collects personal and institutional details, information about their proposed data usage, and information which demonstrates their expertise and ability to conduct the research described in a competent and secure manner. They will also be required, if they have not already done so, to agree and sign the standard UKDA End User License, and also agree/sign any Special License conditions which apply to the resource they wish to access. This application would be first checked for accuracy, sense and completeness by UKDA staff, and then forwarded to the data owners for their access authorisation. Once authorised, users would be informed and requested to sign up for appropriate training (if they have not already been trained). Upon completion of training, users would be granted permission to access the secure data server, either from their own desktop if the owners of the data they wish to access permit, or from their institution's secure data access room. If their institution does not have a secure data access room, the user will have the option of negotiating access from another nearby institution's safe room (the SDS will offer 'matchmaking' introductions, but the specific arrangements must be under the control of the institution hosting the room, as audit trail security will be their responsibility) or coming to the University of Essex, to access the service onsite.

4. Education & Training

It is well known in data security that people are the weakest links. The training and educating of people prevents them from getting a criminal record. The education couples with stricter legislative protections mentioned above can offer another potentially efficient means of improving confidentiality—efficient because the probability of disclosure can be decreased without imposing costs on rule-abiding researchers.

Before becoming an active user of the SDS, users will have to attend a mandatory training session which will focus first on the user's legal and ethical responsibilities within their SDS user license agreement, the mechanics of how to use the SDS, what they can and cannot do in a remote access setting, and the potential of the collaboratory spaces. The second part will focus on statistical disclosure control, assessment of outputs, and analysis aspects of the particular datasets in the SDS.

Access to the SDS will only be granted after users have attended an SDS training session. There will be vetting of data analysis outputs for disclosure issues by SDS staff, to ensure that nothing escapes the secure data setting which could compromise the data security (safe output). One of the purposes of the training is to give researchers the ability to recognise confidential data in order to distinguish it from

⁶ Statistics and Registration Services Act 2007 § 39 (9).

statistical results that are safe to remove from the SDS — in effect, the training removes the ‘reasonable belief’ defence for a disclosure. We believe that penalties will only be an effective deterrent if they are known about, and it should also be clear that we are much more concerned about prevention than punishment.

5. Benefits

Although this system inevitably means an inconvenience to the user in comparison with their accustomed ability to use EUL data on their desktop with all their favourite local software and networked resources, it is the price that users will gladly pay for local access to data which they might otherwise have had to have been seconded to the ONS VML and travelled to ONS sites to access, or simply have been unable to access at all. The SDS will benefit users in;

- ability to work in their own private work areas or in shared areas with other approved researchers
- Access to enhanced, highly sensitive available data storage in tandem with the related metadata through increased capacity and environmental protection
- Possibility of data linkage exercise with using existing data in the UKDA or other administrative data
- collaborative functionality including survey and document library, SPSS/ STATA code library, knowledge repository, disclosure review and technical assistance
- flexibility, access can be restricted to particular users (safe people) and/or particular locations (safe rooms/machines)
- A self-contained secure 'home away from home' service with familiar analytical environments
- Capability for growth and expansion
- A consolidated environment built from the ground up with security and data protection in mind
- Server management processes including auditing, change control, monitoring and alarm notification

6. Data

A variety of data may potentially be available to users within the SDS. We are in ongoing discussions with the owners of key sensitive data resources about how SDS might assist in broadening the use and utility of these important resources, whilst assuring that all legal, moral and security requirements are met. The specifications of data candidates include;

- More detailed variables from existing ESRC-funded data resources
- More previously unavailable detailed variables from government social surveys
- Other government data previously only available in onsite enclaves, or previously unavailable to academic researchers altogether
- Business data which has commercial sensitivity

- Administrative data; the SDS may be able to provide a secure environment for data linkage activities to researchers whose home institutions lack the technological wherewithal to offer it
- Data previously considered too sensitive or potentially disclosive due to its very nature, such as longitudinal data, medical data, etc.

In addition, the service will allow users to bring in less disclosive data from the UKDA standard End User License holdings, upon request.

7. Challenges

The two main goals of the Secure Data Service are; maximizing the utility of microdata for research purposes while protecting the confidentiality of individual respondents. It should be remembered that access to the confidential data is an exception to the non-disclosure rule that must be justified according to the balance of the public good of the research against the risk of a breach of respondent privacy. The following are therefore the desirable qualities of the SDS strategy; maximize data utility while minimizing the disclosure risk. It should be simple, transparent and acceptable to the users.

There are two problems which make achieving the above desired qualities a daunting task; first, there is no consensus on the definition of what is safe data and second, even more contentious is what information loss means. As any effort to implement confidentiality protection is associated with some loss of information. It is important to define a maximum tolerable risk threshold in which we strike balance for both data utility and disclosure risk.

Maintenance of confidentiality needs a consistent and coherent approach. All in all, we have to trust the researchers. For example, how can we prevent against a manual data copying or using photographs for researchers who remotely have access to the disclosive data. It is true that for majority of researchers, data breach happens for access convenience and not out of a malicious intention and surely remotely desktop access to the data would diminish that temptation. However, the possibility of disclosure is always there, the legal framework and training and education may deter users -who are approved researchers afterall - to perpetrate any confidentiality breaches.

8. Evaluation and monitoring of the outputs

The careful user vetting and the most secure analysis environment in the world cannot on its own ensure that data are not disclosed. The missing piece of the data security puzzle is not what goes into the secure data system, but what comes out of it. For the service to be able to meet the security guarantees placed upon it by the data guardians, it is inevitable that it must offer some form of output screening. If an output has been determined to be disclosive, it will be up to the user to determine the best way to render it safe.

The SDS disclosure advisor will divide outputs from SDS into three main categories:

- Safe: No risk / very low risk of disclosure – output will be released promptly
- Uncertain outputs: Low or medium risk of disclosure – output will be considered carefully, with some dialogue with the researcher as necessary, perhaps to collapse categories, remove one or more variables or suppress some cells
- Unsafe: High risk of disclosure – output will be blocked in its current form and won't be released. This is the responsibility of the researcher to produce safe outputs and demonstrate that they are free from the disclosure risks.

Various techniques can be employed – τ -**ARGUS** focuses on an algorithm of controlled rounding and cell perturbation for tabular data.

There are several solutions available to protect the information of the sensitive cells:

- Combining categories of the spanning variables (table redesign). Larger cells tend to protect the information about the individual contributors better.
- Suppression of additional (secondary) cells to prevent the recalculation of the sensitive (primary) cells.

τ -**ARGUS** has been built around the calculation of the optimal set (with respect to the loss of information) of secondary cells. A typical τ -**ARGUS** session will be one in which the users will first be presented with the table containing only the primary unsafe cells. The user can then choose how to protect these cells. This can involve combining categories, equivalent to the global recoding. The result will be an update of the table with fewer unsafe cells if the recoding has worked. The system can solve the remaining unsafe cells by finding secondary cells to protect the primary cells.

9. Summary

The SDS is a secure environment funded by ESRC to provide researcher access to disclosive micro data either from their offices, safe rooms in their institutions or on site at the UKDA. It has two goals: to promote researcher access to sensitive micro data and to protect confidentiality. Its operation is legally framed by the 2007 Statistics Act which makes access to confidential data for statistical purposes possible.

Researcher access to microdata serves the public good both by leveraging existing public investments in data collection, and by ensuring high quality science through the replication of scientific analysis. The SDS provides Approved/Accredited researchers with remote access to microdata using the most secure methods to protect confidentiality. This is achieved by implementing technological security (Citrix gateway), applying statistical protections, enforcing legal requirements, and training researchers. The SDS/UKDA also ensures that valuable data are preserved for the long term by documenting the data using DDI compliant metadata standards. In addition, the SDS aims to engage the research community in using its collaborative data space to share information which enables collaboration among geographically dispersed researchers.

10. References

- Hunderpool, Anco et al. (2009) Handbook on statistical disclosure control version 1.1, ESSnet-Project. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
- Hunderpool, Anco et al. (2009) τ -ARGUS version 3.3 user's manual. ESSnet-Project. <http://neon.vb.cbs.nl/casc/Software/TauManualV3.3.pdf>
- Mulcahy, Timothy M, & John Niesznel (2008) Towards a Secure Data Service at the UK Data Archive. SDS Consultants' Report.
- Ritchie, F (2006) *Disclosure Control of Analytical Outputs*. Mimeo: Office for National Statistics, UK
- Ritchie, F (2007) Disclosure control for regression outputs, Mimeo : Office for National Statistics, UK.
- Ritchie, F (2007) *Statistical Disclosure Control in a Research Environment*. Mimeo: Office for National Statistics, UK
- Wright, Melanie, (2008) Case for Support – Secure Data Service.